THIS OPINION WAS NOT WRITTEN FOR PUBLICATION

The opinion in support of the decision entered today (1) was not written for publication in a law journal and (2) is not binding precedent of the Board.

Paper No. 11

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE BOARD OF PATENT APPEALS
AND INTERFERENCES

Ex parte ROBERT C. PAULSEN, Jr. and MICHAEL J. MARTINO

Appeal No. 2000-0810
Application 08/699,412[1]

ON BRIEF

Before LEE, GARDNER-LANE and MEDLEY, Administrative Patent Judges.

LEE, Administrative Patent Judge.

DECISION ON APPEAL

This is a decision on appeal under 35 U.S.C. § 134 from the examiner's rejection of appellants' claims 1-20.

References relied on by the Examiner

Kucera et al. (Kucera)	4,773,009	Sep. 20, 1988
Ejiri	5,182,708	Jan. 26, 1993

[1] Application for patent filed August 19, 1996. The real party in interest is International Business Machines Corporation.

The Rejection on Appeal

Claims 1-20 stand rejected under 35 U.S.C. § 103 as being unpatentable for obviousness over Kucera and Ejiri.

The Invention

The claimed invention is directed to a computer implemented system and method for identifying the language in which a document was written. In the Background section of the specification, the appellants state:

> In the prior art, for example, when an electronic document was sent across national boundaries, computer system operations were interrupted so that a human being could determine the natural language of a received document before a given operation such as selecting, displaying, printing, and so forth which may be dependent upon the peculiarities of an given natural language. In the context of an internet search, unless the user is multilingual, he is likely to be interested only in the retrieved documents in his native language.
>
> The invention described herein eliminates the need for such human intervention by automatically determining the correct natural language of the computer recorded document.

The independent claims are claims 1, 8, 15 and 16. All dependent claims depend either directly or indirectly from claims 1, 8 and 16. No claim depends from claim 15. Independent claims 1 and 15 are reproduced below:

1. A method for identifying a language in which a computer document is written, comprising the steps of:

comparing a plurality of words from the document to words in a plurality of word tables, each word table associated with and containing a selection of most frequently used words in a respective candidate language;

accumulating a respective count for each candidate language each time one of the plurality of words from the document is present in the associated word table; and

identifying the language of the document as the language associated with the count having the highest value.

15. A system comprising a memory and a processor for identifying a language in which a computer document is written, wherein a plurality of words from the document are compared to words in a plurality of word tables, each word table associated with and containing a selection of most frequently used words in a respective candidate language, a respective weighted count is accumulated for each candidate language each time one of the plurality of words from the document is present in the associated word table, and identifying the language of the document as the language associated with the count having the highest value, the improvement comprising:

the words in each word table are selected based on frequency of occurrence in a candidate language so that each word table covers an equivalent percentage of the associated candidate language.

Discussion

A reversal of the rejection on appeal should not be construed as an affirmative indication that the appellants' claims are patentable over prior art. We address only the positions and rationale as set forth by the examiner and on which the examiner's rejection of the claims on appeal is based.

Three features are expressly recited in each of independent claims 1, 8 and 16: (1) there is a plurality of word tables each of which contains a selection of most frequently used words in a respective candidate language; (2) accumulating a

respective count for each candidate language each time one of the plurality of words from the document at issue is present in the associated word table; and (3) identifying the language of the document as the language associated with the count having the highest value. It is manifestly evident that the reference to "the highest value" concerns the accumulated count for each candidate language, the only "count" previously defined in these claims.

The examiner is correct in noting that the term "language" as is used in the appellants' claims is not limited to natural languages but is sufficiently broad to cover various variations of genres of the same language. Note that the appellants' specification on page 6, in the first paragraph of the Detailed Description of the Drawings section, states:

> In this specification, the term "language" means a natural language, i.e. human language, used for human communications, e.g., English, French, Spanish, German, and so forth. The term "language", as used in the claims, also applies to "genres" within a natural language. Genre is defined herein as a distinctive style of use of a language in some specific context. For example, genre within English includes technical writing, business writing, legal writing, medical writing, fiction, and many others. Thus, genre applies to different variations of the same language involving different styles and manners of word use within a natural language which are reflected in coded documents, and may involve a distinctive use of standard words in a language or may add new words to a language. Genre may reflect particular jargons and dialects of a language such as High German, Low German, and Swiss German, or as London English and New York English.

However, several problems undermine the validity of the examiner's rejection of appellants' independent claims 1-14 and 16-20.

First, the examiner erroneously construed the scope and content of the disclosure of Kucera. The following finding of the examiner (Answer at 8) is incorrect:

> In this case, Kucera's text analyzer utilizes a list of the most frequently used word in the English language to determine readability scores based on the tallying the number of familiar words in a body of text (Kucera; col. 9, lines 43-45 and col. 14, lines 52-60). Since readability is "a measure of the style difficulty of text" (Kucera; col. 1, lines 60-61), and because "genre applies to different variations of the same language involving different styles and manners of word use within a natural language" according to Appellant's specification, **Kucera is directed to mathematically determining the genre of a document** based on the tallying [of] the number of familiar words in a body of text. (Emphasis added.)

The above-quoted rationale is simply that (1) because Kucera discloses use of a list of most frequently used words in the English language to determine a "readability" score for a document, (2) because "readability" is a measure of the style difficulty of text, and (3) because "genre" refers to different variations of the same language involving different styles and manners of word use, Kucera discloses determining the genre of a document based on tallying the number of familiar words in a body of text.

The problem, however, is that Kucera does not disclose the identification of any recognized style or manner of word use. Kucera simply produces a "readability" score based on well-known readability formulas. It does not categorize ranges of "readability" scores into language styles or genres or identify any particular style or groups of styles

based on a document's readability score. Other than that the word "style" is used in the term "style difficulty" which refers to what a "readability score" measures, the examiner has not pointed to any evidence that Kucera identifies or distinguishes any style of writing from among a plurality of writing styles. No basis has been established to equate a readability measurement of "style difficulty" with the identification of any distinctive language style.

"Style difficulty" refers to how hard it is for the document to be read, whatever is the style of the language used. In Kucera's disclosed system, what is important is how readable a document is, not identification of the style of the language in which the document is written. "Readability" of a document may vary depending on the style of the language in which the document is written. Therefore, the style of a language affects the readability of a document. However, it is evident that more than one language style may result in the same or similar level of readability. For instance, very long sentences, very long paragraphs, and/or absence of punctuation, etc. may result in poor readability of a document, regardless of the language or language style. The examiner has not presented evidence that mere "readability" measurement alone identifies a distinctive language style. Also, in the context of these claims, it is implicit that the source language (inclusive of styles within the same language) in which a document is written pre-exists the creation of that document, and does not simply spring into existence when one determines how "readable" a document may be.

Accordingly, it is without adequate basis for the examiner to find that "Kucera is directed to mathematically determining the genre of a document based on the tallying [of] the number of familiar words in a body of text." The scope and content of Kucera has been misstated by the examiner.

Secondly, each of claims 1, 8 and 16 requires a plurality of tables each of which containing a selection of most frequently used words in a respective candidate language. With regard to this feature, the examiner relied on Kucera's teaching of only a single list of frequently used words in the language in which the document is written. The examiner's position is that a mere duplication of parts is well within the basic knowledge and skill possessed by one with ordinary skill in the art. That view is without merit, as applied to the situation here.

The examiner has articulated no reason for one with ordinary skill in the art to extend Kucera's system to having multiple lists of frequently used words, one list for a different candidate language as is required by the appellants' claims. An example of mere duplication of the sort the examiner refers to would be using two shorter lists instead of one long list, or using three or four bolts instead of two bolts to fasten two plates together. Here, the idea of having multiple word tables where each word table is associated with a different candidate language and stores a frequently used word belonging to that candidate language cannot be regarded as a mere duplication of parts, because no structure initially exists in Kucera's system for processing documents in a second candidate language. Where there is nothing to begin with for a second

language, having a list or table in the first instance for a second language is not a mere duplication of what existed before. The examiner has not articulated and established a motivation for one with ordinary skill in the art to expand Kucera's system to support multiple languages and to support them by having a separate word table of most frequently used words for each.

Third, each of claims 1, 8 and 16 calls for accumulating a respective count for each candidate language when there is a match with a word in the corresponding word table, and identifying the language as that associated with the count having the highest value. The appellants argue that these claims require the use of a "raw count" or "non-normalized sum." The examiner disputes that appellants' claims specify identification of the language used on the basis of a raw count or non-normalized value and thus ignores this limitation. But at least with respect to claims 1, 8 and 16 and claims which depend from claims 1, 8 and 16, the highest value of the accumulated counts for each table determines and identifies the language. That means raw count or non-normalized value is indeed used to make the language identification. The examiner's view is without merit and has provided no sufficient basis to refute the raw count or non-normalized value argument of the appellants.

Ejiri is relied on by the examiner for its statement that the frequency of a word or n-gram has been used as a clue to identify an author or language. That general disclosure does not remedy the deficiencies of Kucera regarding the appellants' claimed invention. It does not disclose the use of multiple tables, much less multiple

tables one for each candidate language; and it does not disclose using the highest value of a raw count to determine the language or genre of a language.

Claim 15 is not much different from claims 1, 8 and 16, except that the "raw count" feature of claims 1, 8 and 16 is replaced by a "weighted count." Our discussion above with respect to claims 1, 8 and 16, except for that concerning using the raw count, also applies to claim 15.

Claim 15 additionally recites:

> the words in each word table are selected based on
>
> frequency of occurrence in a candidate language so that each word table covers an equivalent percentage of the associated candidate language.

The examiner concludes, without citation to corresponding teachings in the prior art, that it would have been obvious to one with ordinary skill to select words with respect to each candidate language and maintain an equal percentage of candidate words in each table. The examiner states that the motivation is to not favor one candidate language over another. However, the prior art cited by the examiner does not support the conclusion. On this record, as is argued by the appellants, the motivation of not favoring one language over another by including the same percentage of words in each word table stems only from the appellants' own disclosure. Kucera does not even disclose the use of multiple tables one for each candidate language, let alone suggest to keep the same percentage of frequently used words in each table. The same is true with respect to the Ejiri. The appellants correctly state that the mere fact that a

reference may be modified to result in a claimed invention does not render the modification obvious unless the prior art suggested the desirability of the modification. In re Fritch, 972 F.2d 1260, 23 USPQ2d 1780 (Fed. Cir. 1992).

For the foregoing reasons, the rejection of claims 1-20 under 35 U.S.C. § 103 as being unpatentable over Kucera and Ejiri cannot be sustained.

We commend the examiner's efforts in other parts of the rejection and answer, particularly that explaining, correctly, to the appellants that the test for obviousness is what the combination of prior art teachings would have suggested to one with ordinary skill in the art, not whether the parts of one prior art system may be bodily incorporated into the structure of another without affecting the latter's operation. There is no requirement that the original system of either prior art must be preserved or improved by the results of the proposed combination. That is not the goal. Teachings from a reference are not limited to the preferred embodiments of the reference or even the invention any prior art patent is attempting to protect. Rather, a prior art reference is good for all that it fairly discloses by way of technology, as viewed from the perspective of one with ordinary skill in the art.

Conclusion

The rejection of claims 1-20 under 35 U.S.C. § 103 as being unpatentable for obviousness over Kucera and Ejiri is **reversed**.

REVERSED

JAMESON LEE	)
Administrative Patent Judge	)
	)
	)
	) BOARD OF PATENT
SALLY GARDNER-LANE	) APPEALS AND
Administrative Patent Judge	) INTERFERENCES
	)
	)
	)
SALLY C. MEDLEY	)
Administrative Patent Judge	)

JL:yrt

Attorney for the appellants:

Jeffrey S. Labaw, Esq.
International Business Machines Corporation
11400 Burnet Road
Internal Zip 4054
Austin, Texas 78758